Phương pháp học hiệu quả cho mô hình Biterm Topic Model

Nguyễn Bá Cương

Viện Công nghệ thông tin và Truyền thông
Đại học Bách khoa Hà Nội

Giáo viên hướng dẫn: Ths. Ngô Văn Linh

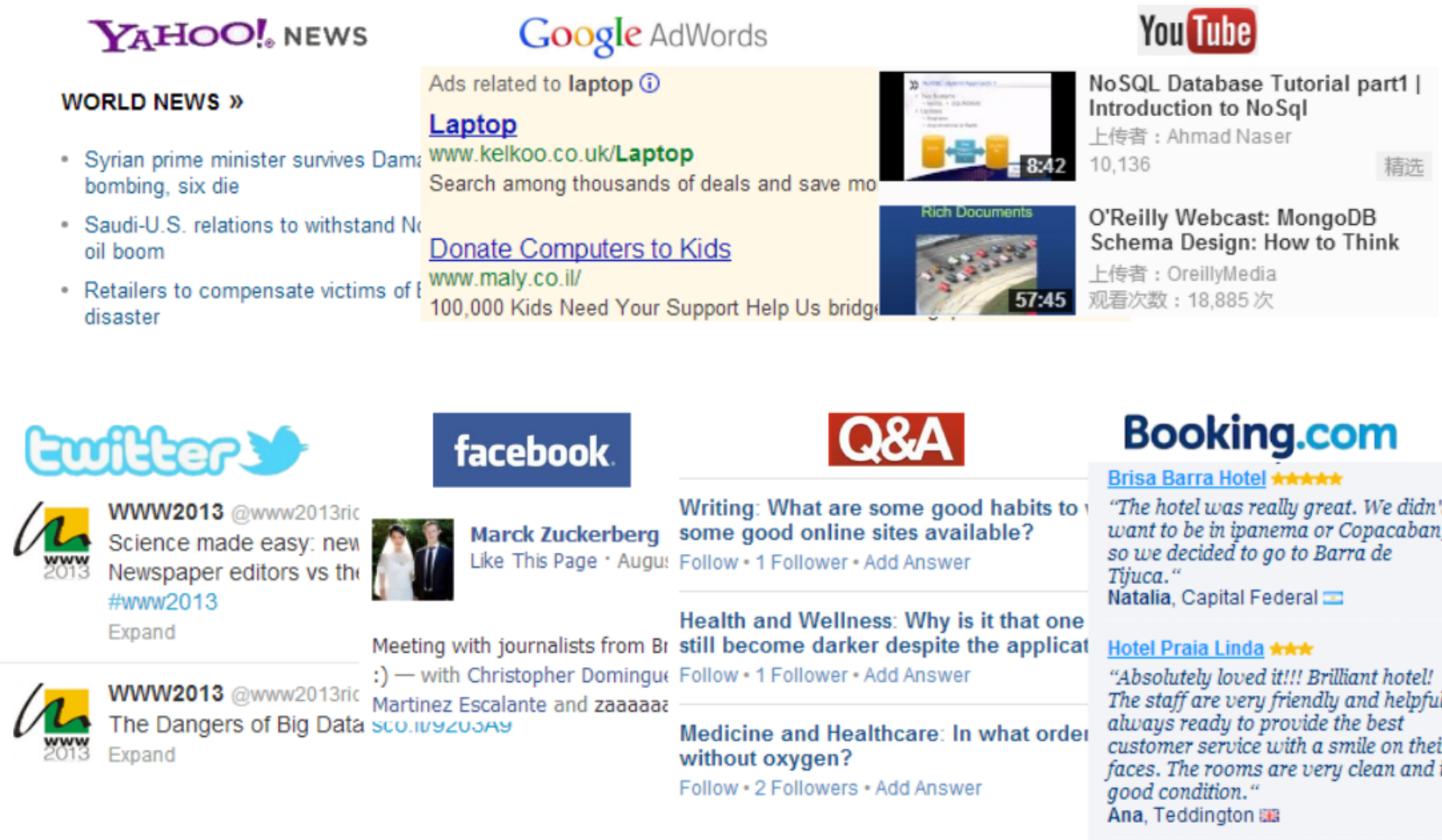
Hà Nội 6, 2017

Nội dung

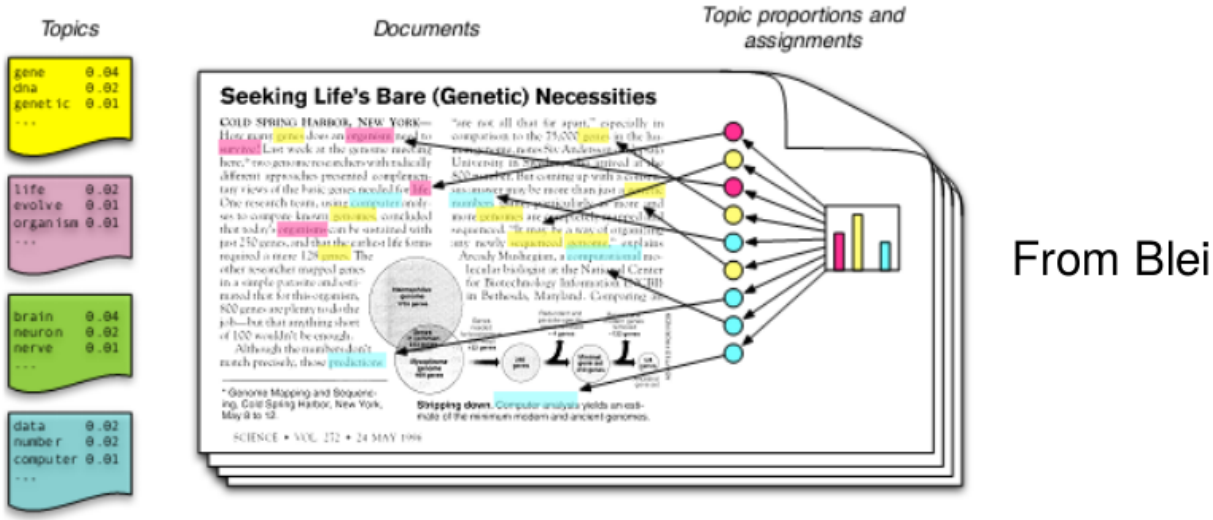
Nội dung

Gới thiệu về short texts

Những văn bản ngắn rất là phổ biến trên các trang web



Mô hình chủ đề



From Blei

(a) Mô hình chủ đề

Mô hình sinh của các văn bản với chủ đề ẩn

- Một chủ đề $\sim$ một phân phối xác suất trong tập từ vựng
- Một văn bản $\sim$ một tập trộn của các chủ đề ẩn
- Một từ $\sim$ một điểm trong một chủ đề

Mô hình chủ đề truyền thống cho short texts

Vấn đề:

- Không đủ ngữ cảnh để xác định ý nghĩa của câu
- Các đặc điểm short texts
  - Độ dài văn bản rất ngắn
  - Số lượng dữ liệu của văn bản ngắn là rất lớn và tăng nhanh
  - Các chủ để nó phản ánh đến các xu hướng xã hội

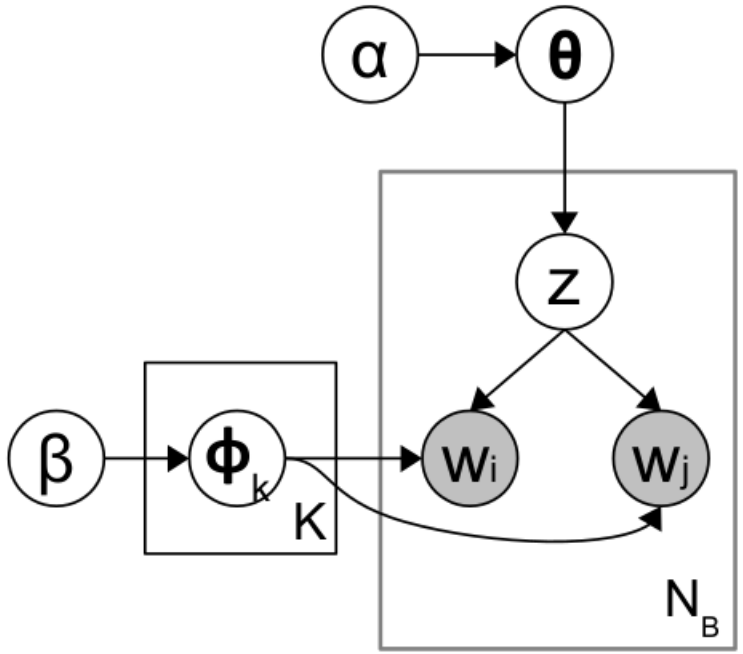
Một số giải pháp:

- Khái thác kiến thức từ bên ngoài để làm giàu sự biểu diễn cho short texts
  $=>$ Việc tìm kiếm dữ liệu từ bên ngoài mất nhiều chi phí.
- Kết hợp một số văn bản ngắn thành một văn bản dài dựa trên một số thông tin. Như tập hợp các bài đăng trên Twitter bởi cùng một người dùng, hay cùng hashtags.
  $=>$ Việc làm này là không mang tính tổng quát

Mô hình Biterm Topic Model (BTM)

Ý tưởng chính

- Chủ đề là tập các từ tương đồng. Các từ cùng xuất hiện trong cùng một văn bản
  => Tại sao không mô hình trực tiếp các từ đồng xuất hiện để học các chủ đề.
- Mô hình chủ đề chịu nhiều vấn đề từ văn bản ngắn
  => Tại sao không sử dụng toàn toàn bộ dữ liệu.



(b) Mô hình BTM

Nội dung

Phương pháp Gibbs Sampling

- Ý tưởng: Sinh ra một mẫu phân bố hậu nghiệm bằng cách duyệt qua các giá trị từ tập dữ liêu ban đầu.
- Vấn đề của phương pháp Gibbs-sampling
  - Không xác định được thời gian hội tụ
  - Dữ liệu biterm sinh ra quá lớn
    => Gibbs-sampling cho thời gian chạy rất lâu.

Phương pháp suy diễn biến phân (VB)

- Thay vì tối ưu hóa trực tiếp hàm mục tiêu, ta xây dựng một cận dưới cho hàm mục tiêu và tối ưu hóa trên cận dưới đó.
- Thuật toán thực hiện hai bước như sau
  - Bước E ta tính giá trị

$$t_{n,k} = \frac{\phi_{k,w_{n,1}} \phi_{k,w_{n,2}} \theta_k}{\sum_k \phi_{k,w_{n,1}} \phi_{k,w_{n,2}} \theta_k}$$

  - Bước M ta cập nhật giá trị

$$\theta_k = \frac{\sum_n t_{n,k} + \alpha}{\sum_{k'} \left(\sum_n t_{n,k'} + \alpha \right)}$$

$$\phi_{k,w} = \frac{\sum_n t_{n,k}\ c(b_n, w) + \beta}{W\beta + \sum_n 2t_{n,k}}$$

Algorithm 1 Thuật toán Online VB cho mô hình BTM

Input: Số lượng chủ đề $K, \alpha, \beta, \{\gamma_i\}_{i=1}^{T}$, tập dữ liệu $B^{(1)}, ..., B^{(T)}$

Output: ϕ, θ

Khởi tạo ngẫu nhiên $\phi, \theta, S^0_{\theta_k} = 0, S^0_{\phi_{k,w}} = 0, i = 0$

for $i = 1$ to ∞ **do**

  for each biterm $b_j = (w_{i,1}, w_{i,2}) \ni B^{(i)}$ **do**

    $t_{j,k} \propto \phi_{k,w_{j,1}} \phi_{k,w_{j,2}} \theta_k$

  end for

  $S^i_{\phi_{k,w}} = \sum_j t_{i,k} c(b_j, w)$; $S^i_{\theta_k} = \sum_j t_{j,k}$

  $S^i_{\phi_{k,w}} = (1 - \gamma_i) S^{(i-1)}_{\phi_{k,w}} + \gamma_i S^i_{\phi_{k,w}}$; $S^i_{\theta_k} = (1 - \gamma_i) S^{(i-1)}_{\theta_k} + \gamma_i S^i_{\theta_k}$

  #Cập nhật tham số mô hình

  $\theta_k \propto S^i_{\theta_k} + \alpha$; $\phi_{k,w} \propto S^i_{\phi_{k,w}} + \beta$

end for

Phương pháp học ngẫu nhiên

- Trong quá trình học, ở mỗi minibatch, chúng ta bỏ đi một phần dữ liệu của minibatch đó.
  - Dữ liệu bỏ đi là một cách ngẫu nhiên. Bỏ đi từng biterm trong tập dữ liệu với xác suất là p.
  - Ở mỗi vòng lặp, dữ liệu bỏ đi không được sử dụng.
- Ý tưởng này áp dụng rộng cho nhiều phương pháp học khác.
  - Áp dụng cho phương pháp VB => RVB
  - Áp dụng cho phương pháp Gibbs-sampling => R-Gibbs-sampling

Algorithm 2 Thuật toán Online RVB cho mô hình BTM

Input: Số lượng chủ đề $K, \alpha, \beta, \{\gamma_i\}_{i=1}^{T}$, tập dữ liệu $B^{(1)}, ..., B^{(T)}, p$

Output: ϕ, θ

1: Khởi tạo ngẫu nhiên $\phi, \theta, S^0_{\theta_k} = 0, S^0_{\phi_{k,w}} = 0, i = 0$

2: **for** $i = 1$ to ∞ **do**

3: Chọn một lượng các biterm trong mỗi mini-batch $B^{(i)}$.
Mỗi biterm trong minibatch thì khả năng bỏ đi là một giá trị xác suất p. Tập dữ liệu sau khi bỏ đi là C

4: **for each** biterm $b_j = (w_{i,1}, w_{i,2}) \ni C$ **do**

5: $t_{j,k} \propto \phi_{k,w_{j,1}} \phi_{k,w_{j,2}} \theta_k$

6: **end for**

7: $S^i_{\phi_{k,w}} = \sum_j t_{i,k} c(b_j, w)$; $S^i_{\theta_k} = \sum_j t_{j,k}$

8: $S^i_{\phi_{k,w}} = (1 - \gamma_i) S^{(i-1)}_{\phi_{k,w}} + \gamma_i S^i_{\phi_{k,w}}$; $S^i_{\theta_k} = (1 - \gamma_i) S^{(i-1)}_{\theta_k} + \gamma_i S^i_{\theta_k}$

9: $\theta_k \propto S^i_{\theta_k} + \alpha$; $\phi_{k,w} \propto S^i_{\phi_{k,w}} + \beta$

10: **end for**

Algorithm 3 Thuật toán Online R-Gibbs-sampling cho mô hình BTM

Input: K, α, β, tập dữ liệu $B^{(1)}, ..., B^{(T)}, p$
Output: $\{\phi^{(t)}, \theta^{(t)}\}_{t=1}^{T}$

Khởi tạo $\alpha^{(1)} = (\alpha, ..., \alpha)$ và $\{\beta_k^{(1)} = (\beta, ..., \beta)\}_{k=1}^{K}$
for $t = 1$ to T **do**
  Khởi tạo ngẫu nhiên các giá trị chủ đề cho tất cả các biterm
  for $iter = 1$ to N_{iter} **do**
    Chọn một lượng các biterm trong mỗi mini-batch $B^{(t)}$.
    Mỗi biterm trong minibatch thì khả năng bỏ đi là một giá trị xác suất p. Tập dữ liệu sau khi bỏ đi là C
    for each biterm $b_i = (w_{i,1}, w_{i,2}) \ni C$ **do**
      Tính giá trị k và cập nhật $n_k^{(t)}, n_{w_{i,1}|k}^{(t)}$ và $n_{w_{i,2}|k}^{(t)}$ theo k
    end for
    Cập nhật $\alpha^{(t+1)}$ và $\{\beta_k^{(t+1)}\}_{k=1}^{K}$
  end for
  Cập nhật $\phi^{(t)}$ và $\theta^{(t)}$
end for

Những điểm mạnh của phương pháp học mới

Phương pháp VB:

- Thời gian học của phương pháp VB nhanh hơn so với phương pháp *Gibbs-sampling* mà tác giả đề xuất.
- *VB* cho chất lượng mô hình cao hơn so với phương pháp học bằng *Gibbs-sampling*

Phương pháp học ngẫu nhiên:

- Phương pháp học ngẫu nhiên cho thời gian chạy nhanh hơn so với phương pháp gốc thực hiện.
- Phương pháp học ngẫu nhiên cho chất lượng chủ đề xấp xỉ như kết quả mà phương pháp gốc thực hiện

Nội dung

Bộ dữ liệu thử nghiệm

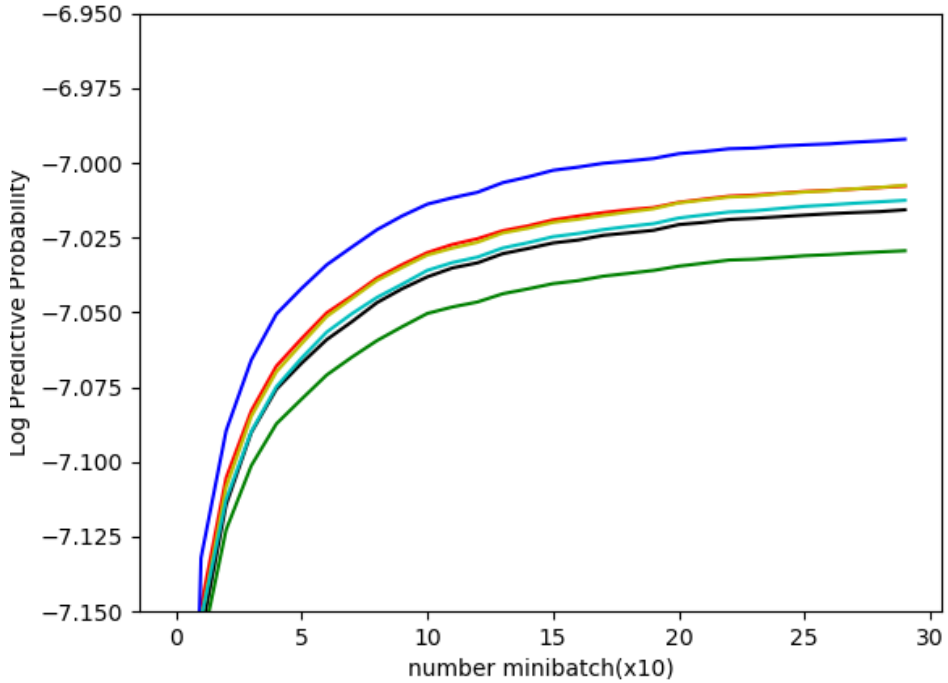
	Số lương văn bản	Độ dài trung bình	V
Twitter	1,485,068	10.14	89,474
Yahoo Questions	537,770	4.73	24,420
Nytimes Titles	1,684,127	5.15	55,488

Bảng : Bảng mô tả dữ liệu thử nghiệm

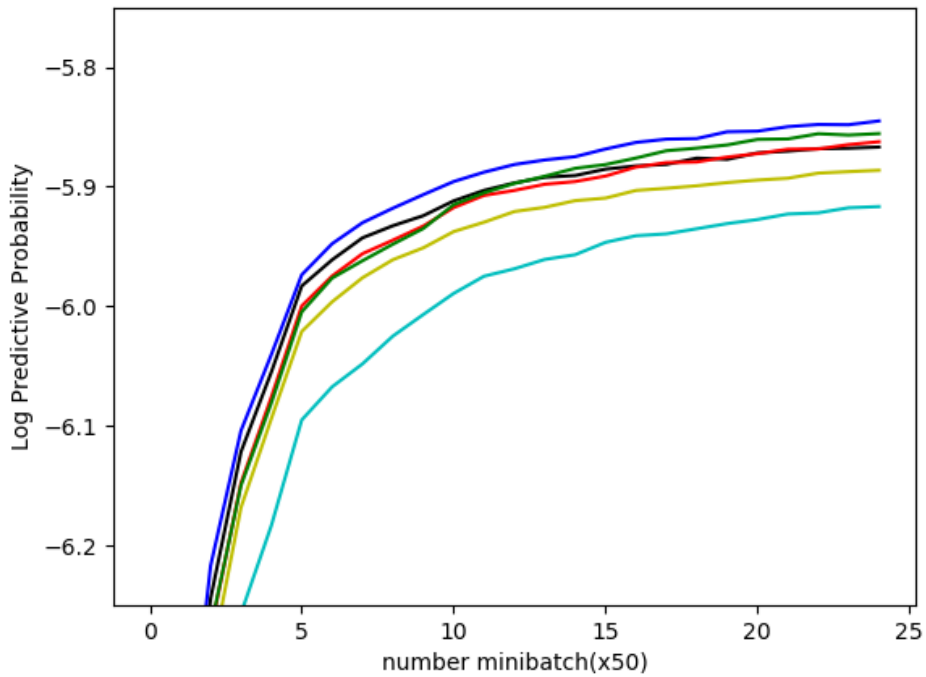
Tiêu chí đánh giá

- Thử nghiệm so sánh trên 3 phương pháp:
  - Gibbs-sampling
  - Suy diễn biến phân - VB
  - Học ngẫu nhiên: Phương pháp RVB và R-Gibbs-sampling
- Tiêu chí đánh giá:
  - Khả năng phán đoán mô hình: Sử dụng độ đo *log predictive probability*
  - Chất lượng chủ đề: Sử dụng độ đo NPMI
  - Thời gian chạy

Độ đo khả năng phán đoán mô hình



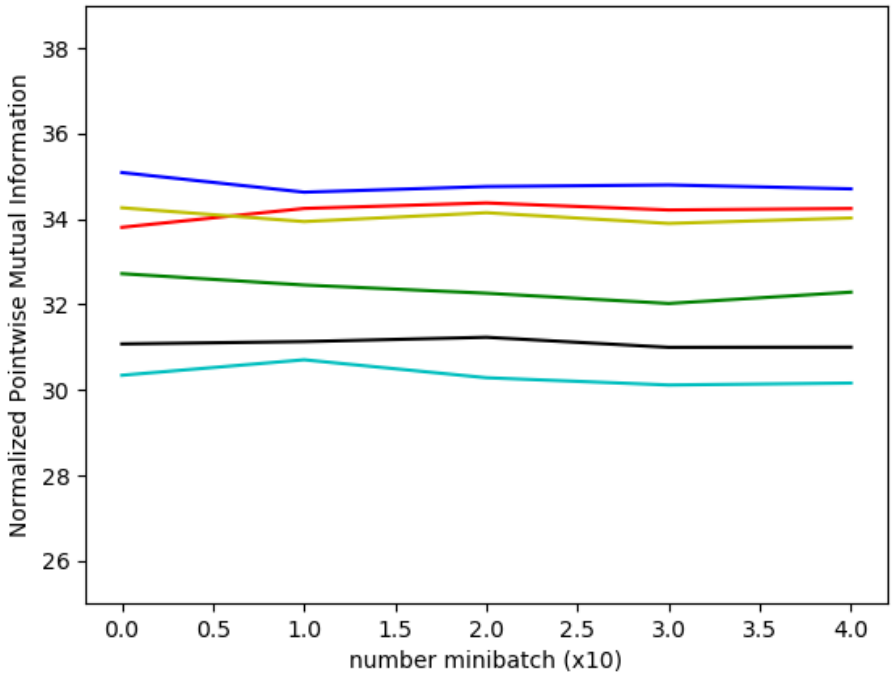
(c) Online Gibbs-sampling và R-Gibbs-sampling



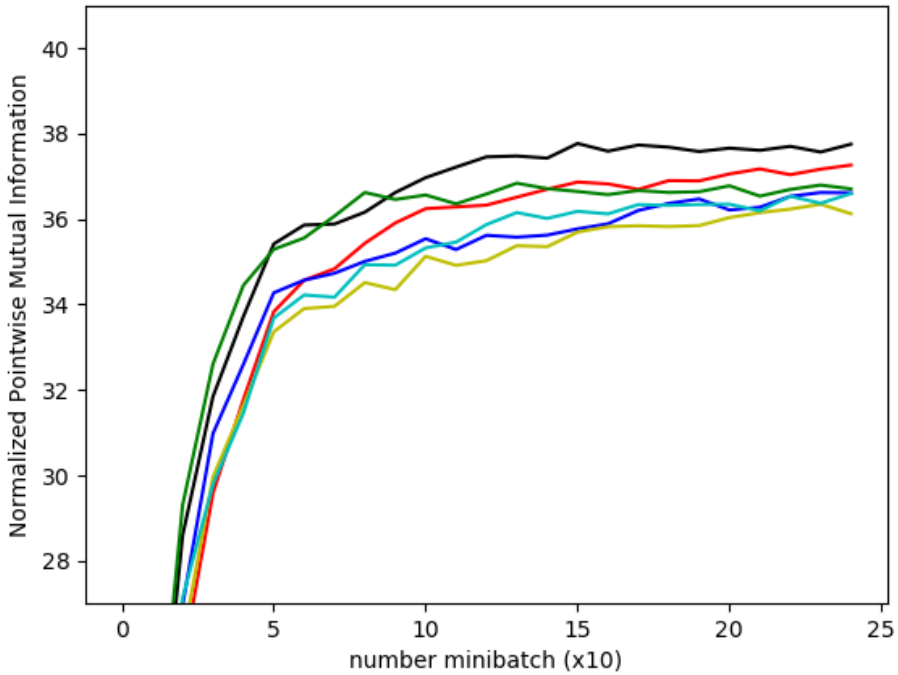
(d) Online VB và RVB

Hình : Kết quả độ đo khả năng phán đoán mô hình cho bộ dữ liệu Twitter

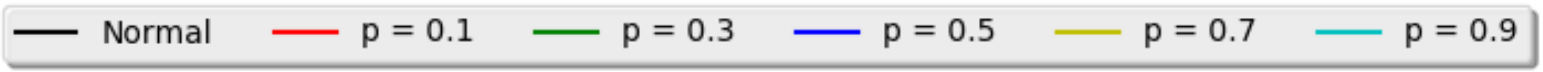
Độ đo chất lượng chủ đề



(a) Online Gibbs-sampling và R-Gibbs-sampling



(b) Online VB và RVB



Hình : Kết quả độ đo chất lượng chủ đề cho bộ dữ liệu Twitter

Thời gian huấn luyện

	K	origin	p = 0.1	p = 0.3	p = 0.5	p = 0.7	p = 0.9
Gibbs	50	249935	231575	186822	80132	56739	24534
VB	50	3592	3463	2922	2166	1249	375
Gibbs	100	404573	357891	288019	213525	134510	46840
VB	100	3224	3050	2503	1870	1189	880
Gibbs	150	460189	429685	350120	297817	296984	114108
VB	150	3866	3696	3052	2429	1536	781
Gibbs	200	539904	538529	456806	387955	275327	171603
VB	200	4925	4594	3313	2622	1944	1322

Đánh giá

- Phương pháp Gibbs sampling, cho thời gian chạy rất chậm.
- Phương pháp VB cho thời học vượt trội hơn hẳn, và kết quả chất lượng mô hình cao hơn so với phương pháp học Gibbs sampling.
- Phương pháp học ngẫu nhiên cho thời gian học nhanh hơn nhiều lần ứng với lượng dữ liệu được bỏ đi.

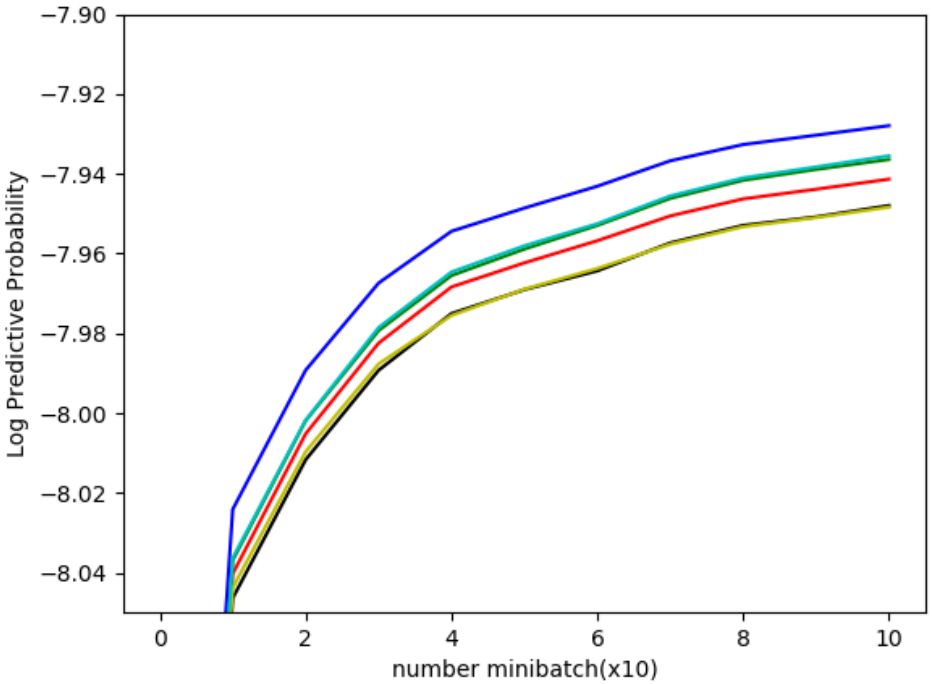
Nội dung

Kết luận

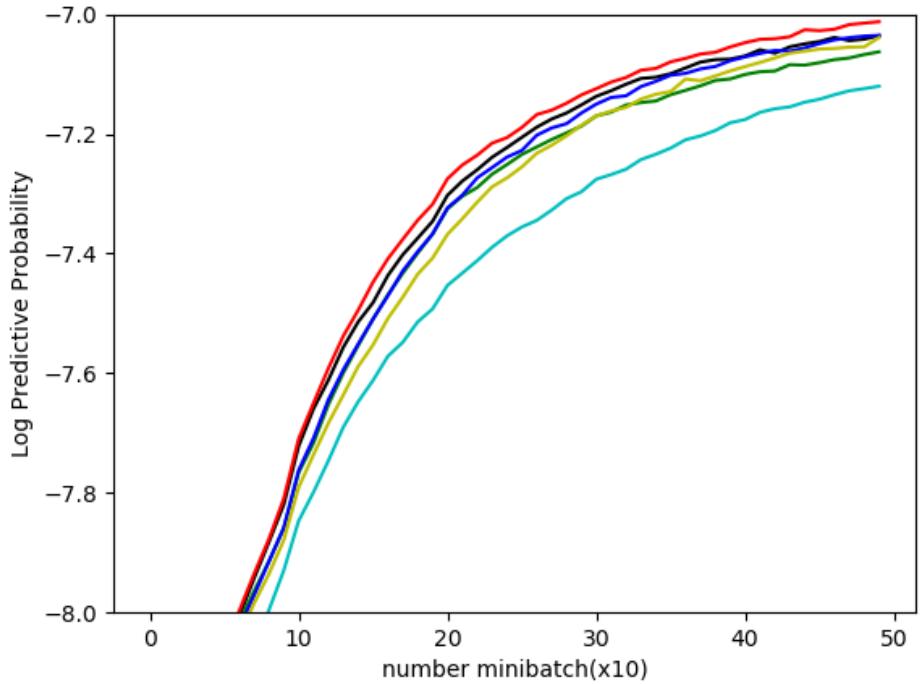
- Phương pháp học VB
  - Kết quả cho thấy chất lượng chủ đề hay khả năng phán đoán của mô hình tốt hơn hẳn so với phương pháp Gibbs sampling
  - Thời gian học của phương pháp VB vượt trội hơn hẳn so với phương pháp Gibbs-sampling
- Phương pháp học ngẫu nhiên
  - Chất lượng mô hình xấp xỉ với phương pháp gốc.
  - Thời gian học nhanh hơn rất nhiều so với phương pháp gốc.


Thank you!

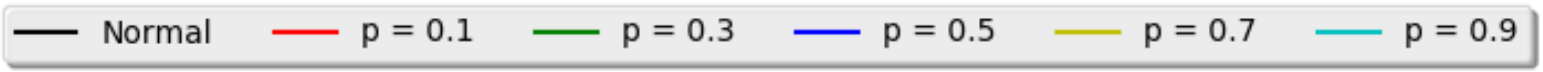
Độ đo khả năng phán đoán mô hình



(a) Online Gibbs-sampling và R-Gibbs-sampling

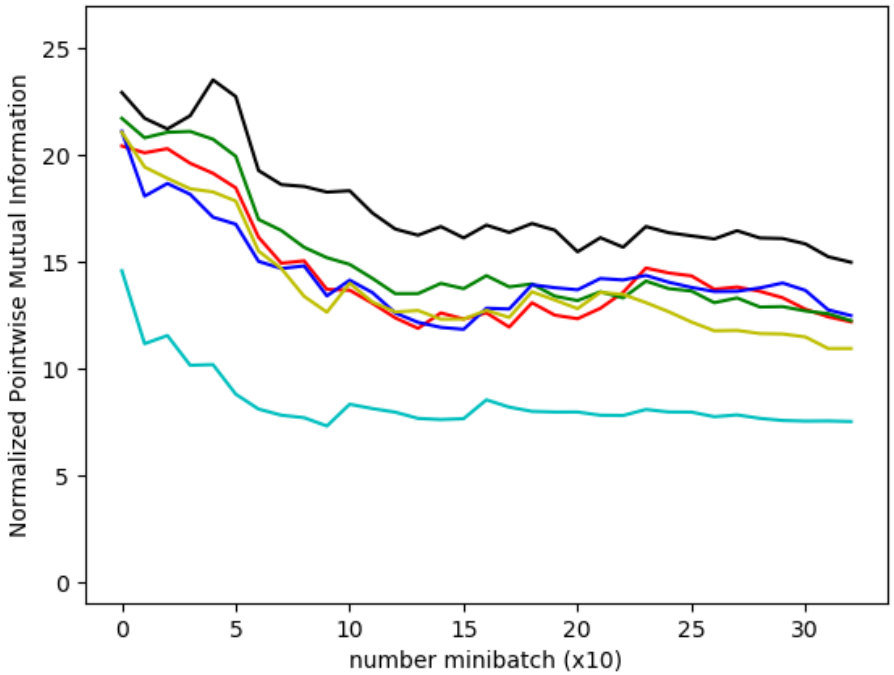


(b) Online VB và RVB

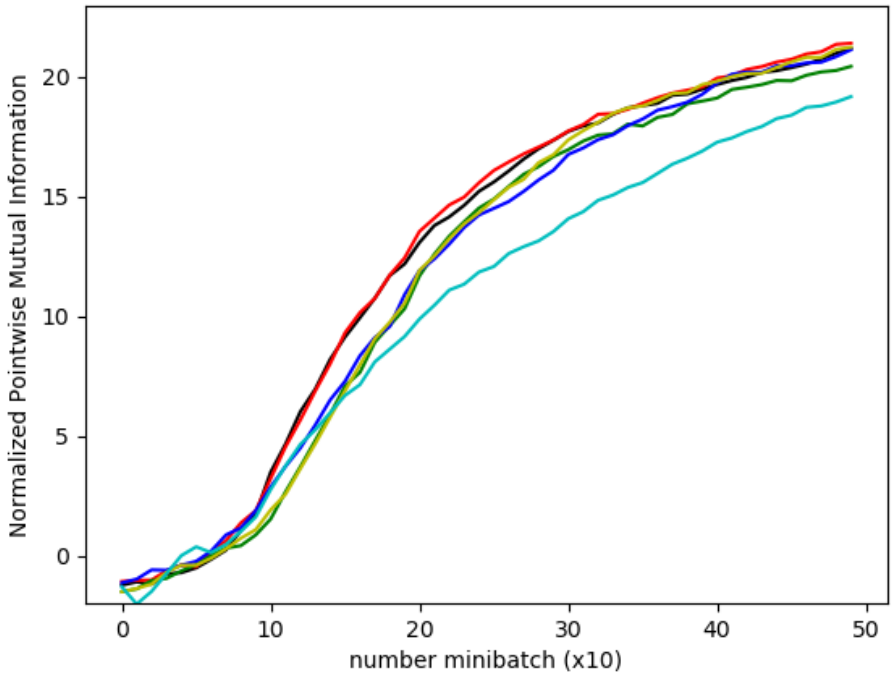


Hình : Kết quả độ đo khả năng phán đoán mô hình cho bộ dữ liệu Yahoo

Độ đo chất lượng chủ đề



(a) Online Gibbs-sampling và R-Gibbs-sampling



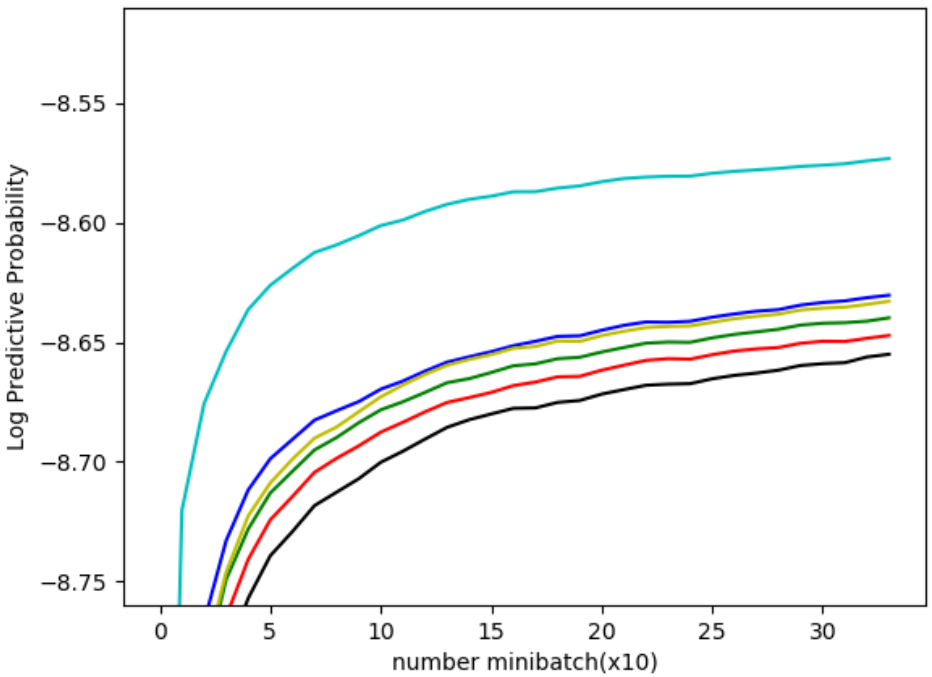
(b) Online VB và RVB

Hình : Kết quả độ đo chất lượng chủ đề cho bộ dữ liệu Twitter

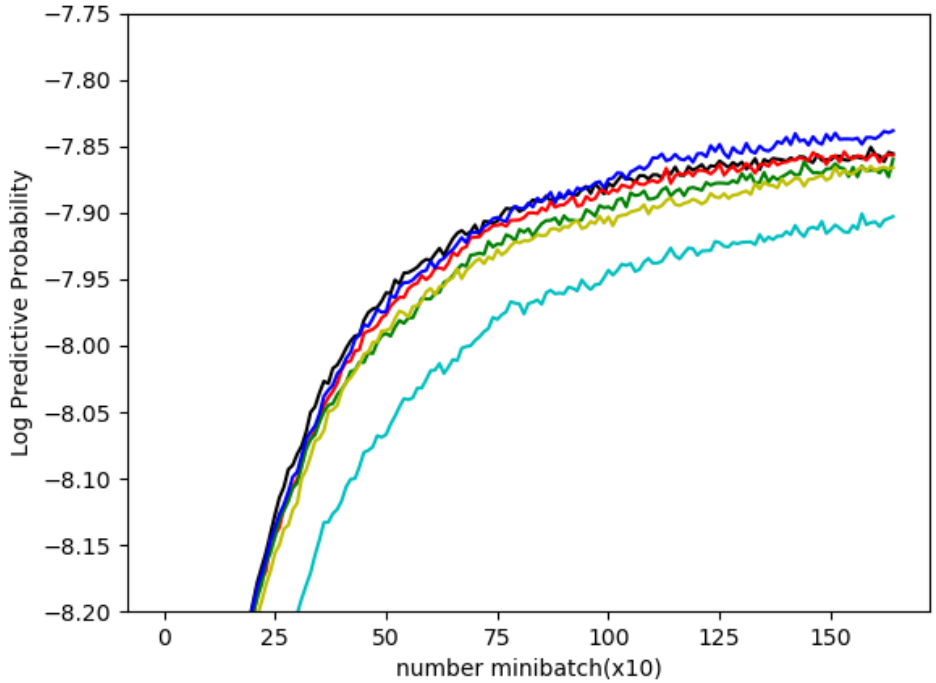
Thời gian chạy với bộ dữ liệu Yahoo

	K	origin	r = 0.1	r = 0.3	r = 0.5	r = 0.7	r = 0.9
Gibbs	50	24259	22605	18919	14154	9594	3856
VB	50	546	534	501	432	388	176
Gibbs	100	32042	27544	21763	16729	19533	7565
VB	100	448	418	372	328	274	223
Gibbs	150	73385	68762	58344	44619	30748	12387
VB	150	514	490	453	398	337	265
Gibbs	200	64068	51862	46133	6406	41653	8894
VB	200	580	559	485	434	373	306

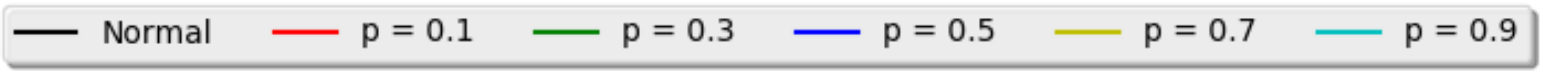
Độ đo khả năng phán đoán mô hình



(a) Online Gibbs-sampling và R-Gibbs-sampling

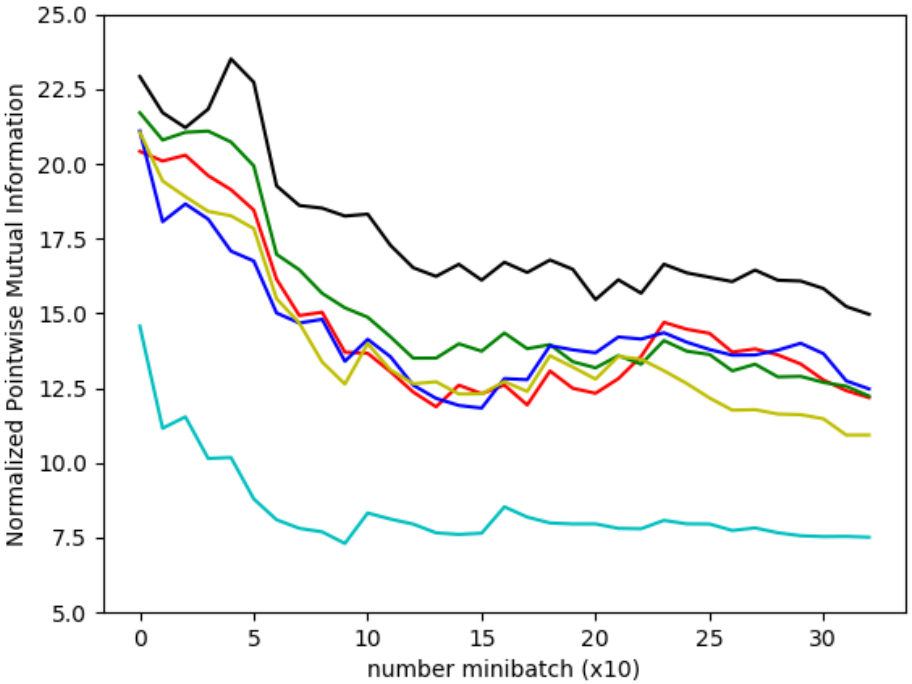


(b) Online VB và RVB

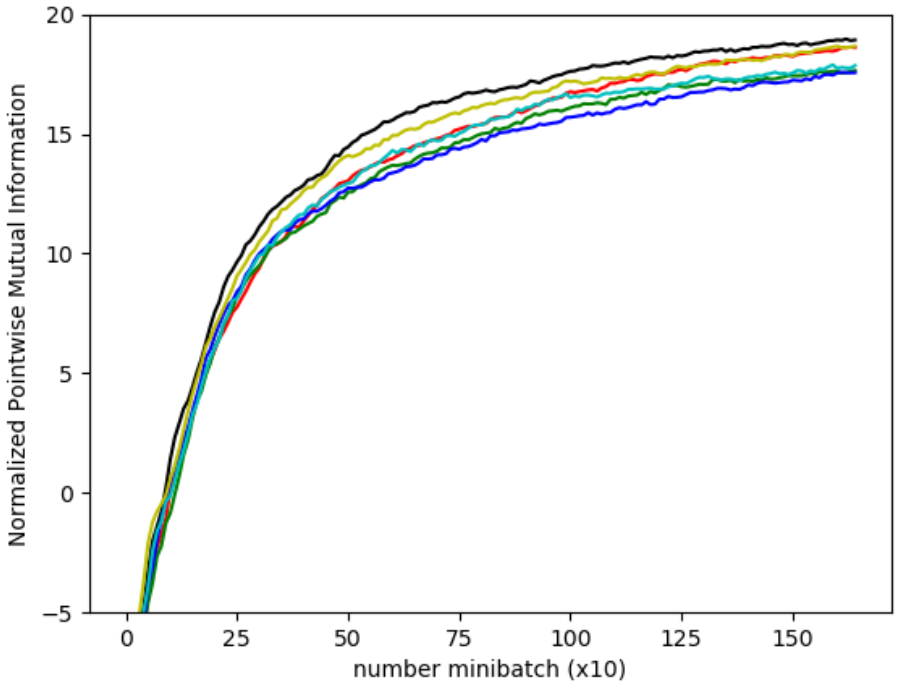


Hình : Kết quả độ đo khả năng phán đoán mô hình cho bộ dữ liệu NYT

Độ đo chất lượng chủ đề



(a) Online Gibbs-sampling và R-Gibbs-sampling



(b) Online VB và RVB

Hình : Kết quả độ đo chất lượng chủ đề cho bộ dữ liệu NYT

Thời gian chạy với bộ dữ liệu NYT

	K	origin	r = 0.1	r = 0.3	r = 0.5	r = 0.7	r = 0.9
Gibbs	50	81029	75909	63424	49757	33146	11787
VB	50	1190	1876	1723	1506	1355	633
Gibbs	100	90219	85610	71573	56306	39515	15900
VB	100	1701	1569	1398	1229	996	880
Gibbs	150	130083	123790	124131	107083	89761	38267
VB	150	1800	1792	1660	1499	1250	1030
Gibbs	200	193110	180111	184658	151754	47640	33396
VB	200	2308	2171	1707	1554	1361	1096